Speech Communication 138 (2022) 80–87

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic syllable stress detection under non-parallel label and data condition

Chiranjeevi Yarra [a,*], Prasanta Kumar Ghosh [b]

[a] Language Technologies Research Center, International Institute of Information Technology (IIIT), Hyderabad, 500032, India
[b] Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, 560012, India

ARTICLE INFO

Keywords:
Stress detection
Non-parallel label and data
Stress label assignment
Stress detection for estimated data

ABSTRACT

Typically, automatic syllable stress detection is posed as a supervised classification problem, for which, a classifier is trained using manually annotated (existing) syllable data and stress labels. However, in real testing scenarios, syllable data is estimated since manual annotation is not possible. Further, the estimation process could result in a mismatch between the lengths of the estimated and the existing syllable data causing no one-to-one correspondence between the estimated syllable data and the existing labels. Hence, the existing labels and estimated syllable data together cannot be used to train the classifier. This can be avoided by manually labeling the estimated syllable data, which, however, is impractical. In contrast, we, in this work, propose a method to obtain labels for estimated syllable data without using manual annotation. The proposed method considers a weighted version of the well-known Wagner–Fisher algorithm (WFA) to assign the existing labels to the estimated syllable data, where the weights are computed based on a set of three constraints defined in the proposed algorithm. Experiments on ISLE corpus show that the performance obtained on the test set for four different types of estimated syllable data are higher when the assigned labels and estimated syllable data are used for training compared to those when existing labels and existing syllable data are used. Also, the label assignment accuracy using the proposed method is found to be higher than that using a baseline scheme based on WFA.

1. Introduction

Automatic detection of syllable stress has been shown to be useful for evaluating pronunciation (Chandel et al., 2007; Zhao et al., 2011; Verma et al., 2006) in several applications including computer assisted language learning (CALL). It is also useful in providing feedback to the second language (L2) learners by automatically identifying localized pronunciation errors (Ferrer et al., 2015; Tepperman and Narayanan, 2005). Typically, the stress detection task is performed as a classification problem in a supervised manner (Tamburini, 2003; Tepperman and Narayanan, 2005; Verma et al., 2006; Deshmukh and Verma, 2009) using a set of features representing a syllable and the respective stress labels (stressed and unstressed). In most of the existing works, the labels are obtained from a manual annotation process and the syllable data (both the syllable transcriptions and their time-aligned boundaries) is estimated using forced-alignment (Ferrer et al., 2015; Shahin et al., 2016, 2014; Deshmukh and Verma, 2009; Li et al., 2011). Hence, the reliability of the stress detection task depends on the features and the quality of the syllable data. In few works, manually corrected syllable data has been used to reduce errors due to forced-alignment (Tepperman and Narayanan, 2005; Shahin et al., 2016). However, in real testing scenarios, the syllable data is obtained mostly using forced-alignment process, where no manual correction is possible. The model used in the forced-alignment determines its accuracy. The accuracy of a forced-alignment scheme, in turn, determines its suitability for obtaining syllable data.

Forced-alignment is performed either with traditional Gaussian mixture model based hidden Markov models (GMM-HMM) (Tepperman and Narayanan, 2005; Shahin et al., 2016; Tamburini, 2003) or with recently proposed deep neural network based HMM (DNN-HMM) acoustic models (Povey et al., 2011), which yield more accurate forced-alignment results. When the speech data from non-native speakers are considered, the reliability of the acoustic models depends on the pronunciation lexicon used in the automatic speech recognition (ASR) training. Further, the lexicon also plays a critical role in the forced-alignment process. A pronunciation lexicon containing all pronunciation variants of L2 learners could result in a better quality syllable data from the forced-alignment. However, availability of such lexicons is limited and identification of such pronunciations is also challenging. Thus, different combinations of acoustic model and lexicons could

* Corresponding author.
E-mail address: chiranjeevi.yarra@iiit.ac.in (C. Yarra).

https://doi.org/10.1016/j.specom.2022.02.001
Received 6 February 2019; Received in revised form 1 November 2021; Accepted 1 February 2022
Available online 22 February 2022

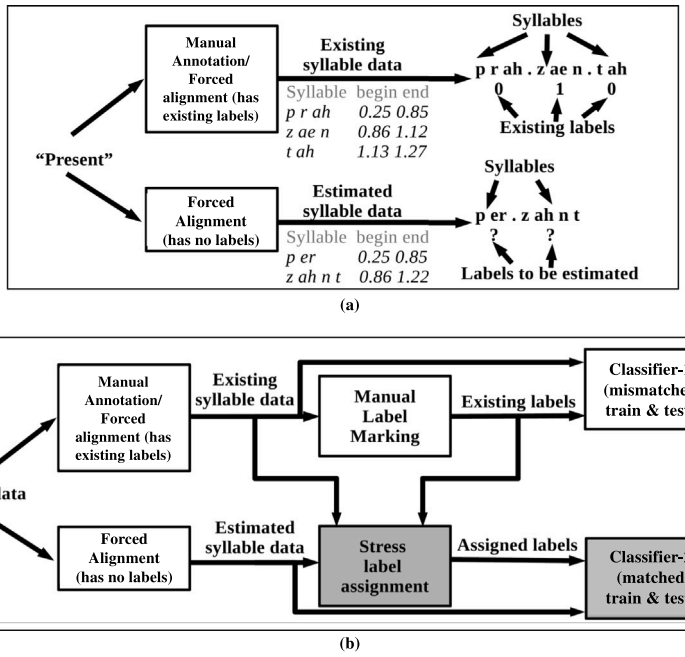

Fig. 1. An illustration of (a) non-parallel data and label scenario for an exemplary word "Present" and (b) mismatched and matched train & test scenarios.

cause variations in the quality of the syllable data in terms of the forced-aligned phonetic transcriptions as well as their boundaries. In addition, a syllable data with readily available labels, in general obtained using manual annotation process, might not be preferable when a better quality forced-aligned syllable data is obtained. As ASR for non-native speech continues to become more accurate with better acoustic model and/or lexicon, the quality of the syllable data from forced-alignment using a state-of-the-art ASR would be better than that obtained with an ASR in the past. Further, it is to be noted that the number of syllables in a word could also vary among the syllable data obtained from different forced-alignment processes.

Hence, available manual (existing) labels for one set of syllable data, referred to as existing syllable data (either manually transcribed or estimated from forced-alignment process), cannot be mapped directly to another set of syllable data, referred to as estimated syllable data, for which labels have to be estimated. This could be because the existing labels do not have one-to-one correspondence with the estimated syllable data, referred to as non-parallel data and label condition. For example, among different models and lexicons used for forced-alignment of the utterances in the ISLE corpus (Menzel et al., 2000), we observe that up to 19.92% of the poly-syllabic words have different number of syllables compared to those available in the corpus. Fig. 1a shows exemplary syllable transcriptions of a spoken word "Present" from existing and estimated syllable data. In the figure, we also indicate existing labels on the syllables from the existing syllable data. From the figure, it is observed that there is mismatch between the existing and estimated syllables in terms of number of syllables as well as their transcriptions. This suggests that the label identification for the estimated syllables is a non-trivial task.

Under this non-parallel condition, a classifier could not be trained with existing labels and estimated syllable data for the stress detection task. However, it is possible to estimate the stress labels on a test set using estimated syllable data and a classifier (Classifier-1 in Fig. 1b) trained with existing syllable data and existing labels. But, the mismatch in the data could degrade the stress detection performance under estimated syllable data conditions in the test phase. In order to avoid this, it is necessary to repeat the manual annotation to obtain stress labels for the estimated syllable data, which is cumbersome, time-consuming and also impractical for different sets of syllable data from different forced-alignment conditions. To circumvent these problems, in this work, we propose a stress label assignment method to assign the existing labels to an estimated syllable data. We hypothesize that the classifier (Classifier-2 in Fig. 1b) trained with the assigned labels and estimated syllable data would perform better on estimated syllable data in the test phase. This is because the proposed method establishes matched train and test conditions by training the classifier with the respective estimated syllable data for which testing is performed.

We perform automatic syllable stress detection using the syllable data obtained from four different combinations of acoustic models and lexicons in the forced-alignment set-up. We obtain stress labels for the estimated syllable data from the existing labels in two steps. In the first step, we pair each phoneme in the existing syllable data with either a phoneme in the estimated syllable data or empty string and vice versa. In order to obtain the pairs, we formulate a string matching problem by defining the constraints similar to the rules used in stress label annotation (Deshmukh and Verma, 2009; Menzel et al., 2000). For this purpose, based on Wagner–Fisher algorithm (WFA) (Wagner and Fischer, 1974), we propose a weighted Wagner–Fisher algorithm (wWFA), in which, weights are computed based on the constraints considered in this work. In the second step, we map the existing labels on the phonemes belonging to syllable nuclei in the existing syllable data to the respective paired phonemes in the estimated syllable data. We use a support vector machine (SVM) classifier for the stress detection using the assigned stress labels and acoustic features (AFs) computed using the work proposed by Yarra et al. (2017).

Experiments are performed on ISLE (Menzel et al., 2000) corpus containing polysyllabic words separately from German and Italian non-native speakers. Stress detection results show that the proposed approach on the estimated syllable data under all four forced-alignment set-ups yields better performance compared to that under the mismatched train-test scenario where existing labels and AFs from the existing syllable data are used for training. Further, in order to know the effect of label assignment on the stress detection, we conduct the experiments on entire data from ISLE corpus, for which manually annotated ground truth stress labels are available for the estimated syllable data from one (among four) forced-alignment set-up considered in this work. The proposed wWFA based label assignment achieves 100% accuracy, while a baseline scheme based on WFA achieves a lesser accuracy of 97.2%.

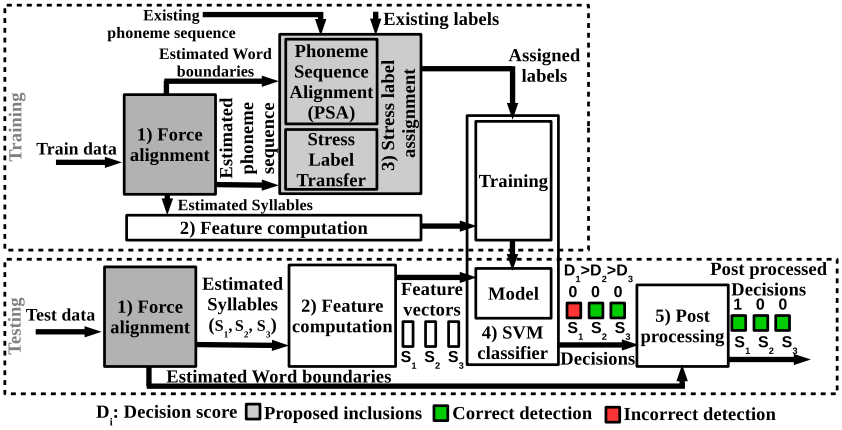

Fig. 2. Block diagram illustrating the steps involved in the proposed approach using a three syllabic word. The gray colored rectangular blocks indicate the proposed inclusions to perform stress detection task in real testing scenarios by estimating the syllable data. The white colored rectangular blocks indicate the components that are considered from the work in the literature (Yarra et al., 2017).

The rest of the paper is organized as follows: Section 2 discusses the proposed approach including stress label assignment, constraints involving stress labeling, proposed wWFA, weights computation based on the constraints, Section 3 describes the corpus details, Section 4 includes the experimental setup, results and discussion on both the stress label assignment and stress detection tasks. The conclusions are summarized in Section 5.

2. Proposed approach

Block diagram in Fig. 2 shows training and testing stages involved in the proposed approach for stress detection task. There are a total of five steps. The first and the second steps are identical for both the stages. In the first step, forced-alignment is done on a speech signal using its word transcription to estimate phoneme transcriptions as well as aligned phoneme and word boundaries. Further, we syllabify the phoneme transcriptions, from which we obtain syllable transcriptions and its time aligned boundaries, referred to as estimated syllable data. In the second step, AFs proposed by Yarra et al. (2017) are computed for each syllable in the estimated syllable data. In the third step, we perform label assignment and map the existing labels to each syllable segment using existing and estimated phoneme transcription as well as word boundaries from forced-alignment. In the fourth step, SVM is trained in the training stage using assigned stress labels & AFs in order to obtain a model for classification. In the testing stage, we classify each syllable segment as stressed or unstressed with the trained SVM model using AFs. In the last step, decision scores from SVM classifier are used to post-process the estimated stress markings to ensure that each polysyllabic word has only one stressed syllable.

Stress label assignment involves two sub-steps – (1) Phoneme sequence alignment (PSA), and (2) Stress label transfer. In the PSA, the existing phoneme sequence (A) is aligned with an estimated phoneme sequence (B) obtained from forced-alignment. In the label transfer, we map the existing stress labels on phonemes indicating syllable nuclei in A to those in B that are paired with the respective nuclei in A from the PSA. We describe these steps in detail in Sections 2.1 and 2.2 respectively.

2.1. Phoneme sequence alignment (PSA)

For the PSA, we propose a string matching algorithm, from which, each phoneme in A paired with either a phoneme in B or empty string (Λ) and vice-versa. The pairing between the sequences A and B is denoted using a trace (T) from A to B, where, in its graphical representation, paired phonemes are connected using a straight line and the remaining phonemes are left unconnected. In general, there are many possible traces from A to B. Fig. 3 shows two exemplary traces between the phoneme sequences A = {f, ao, r} and B = {f, r, er}. The trace in Fig. 3a shows that the phonemes ‘f, ao, r’ in A are paired with the phonemes ‘f, r, er’ in B respectively and are indicated as follows: f → f, ao → r and r → er. Similarly, the trace in Fig. 3b shows that the phonemes ‘f, ao’ in A are paired with the phonemes ‘f, er’ in B and the phoneme ‘r’ in both A and B is left unconnected and these are indicated as follows: f → f, Λ → r, ao → er and r → Λ. Empty string Λ is used in defining the trace because, in a typical string matching, cross-over of these lines are not allowed.

2.2. Stress label transfer

In general, among all the phonemes in a syllable, stress is mainly captured by its nuclei and typically only one syllable nuclei in a word is primarily stressed, referred to as primary stress. The data in ISLE corpus, used in this work, was labeled accordingly. In order to map the existing stress labels on phonemes indicating syllable nuclei in A to the phonemes of syllable nuclei in B, we consider only syllable nuclei pairs among all the paired phonemes between A and B obtained from the PSA. In case a syllable nucleus in B is not paired with a syllable nucleus in A, we propose to consider its label as unstressed. For example, in Fig. 3, the syllable nuclei are ‘ao’ and ‘er’ in A and B respectively and their respective ground-truth labels are stressed (marked with ‘1’). From the trace in Fig. 3a, it is observed that the stress label on ‘ao’ cannot be transferred to ‘er’ directly since it is not paired with ‘ao’. Thus, the stress label on ‘er’ is proposed as unstressed (marked with ‘0’), which does not match with its ground-truth. However, considering the trace in Fig. 3b, the stress label on ‘ao’ is transferred to ‘er’ as ao → er, which results in a label matched with its ground-truth.

Hence, in the PSA, it is necessary to ensure that the label transfer happens across the nuclei in A and B. Also, the transfer should ensure only one primary stress within the phoneme sub-sequence in B corresponding to one word. Therefore, we propose the following constraints to a typical string matching algorithm to obtain PSA.

1. All the stress labels within a sub-sequence in A representing a word must not cross the boundaries of a sub-sequence in B representing the same word, because, the labels are typically assigned by comparing prominence of the syllables belonging to each word (Deshmukh and Verma, 2009; Menzel et al., 2000).
2. At least N syllable nuclei in A and B must be paired, where N is the smaller number of syllables between A and B.
3. The syllable nucleus belonging to a primary stress label of a sub-sequence in A representing a word must be paired with a syllable nucleus of a sub-sequence in B belonging to the same word. This ensures only one primary stress label for a word following PSA.

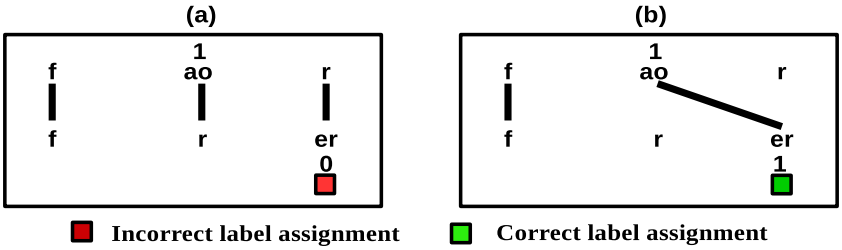

Fig. 3. Two exemplary traces between the phoneme sequences {f, ao, r} and {f, r, er} for the word "for", in which, (a) trace results in an incorrect stress label assignment (b) Expected trace that could result in correct label assignment after the PSA and the stress label transfer.

In this work, for the PSA, we consider Wagner–Fisher algorithm (WFA) (Wagner and Fischer, 1974) and propose modifications to the algorithm, referred to as weighted WFA (wWFA), so that it can incorporate the above constraints. Below, we discuss the steps of the proposed wWFA including the computation of its weights.

2.3. Wagner–Fisher algorithm

This algorithm yields the best alignment between two sequences $A = A(1:n) = \{a_1, a_2, \ldots, a_n\}$ and $B = B(1:m) = \{b_1, b_2, \ldots, b_m\}$ by minimizing the cost C_T given as:

$$C_T(A,B) = \sum_{(i,j)\in T} \gamma(a_i \to b_j) + \sum_{i\in I} \gamma(a_i \to \Lambda) + \sum_{j\in J} \gamma(\Lambda \to b_j) \quad (1)$$

where $a_i, 1 \le i \le n$ & $b_j, 1 \le j \le m$ are the phonemes in the respective sequences A & B. T denotes a trace from A to B. It has been shown that the trace is sufficient to know the alignment between A and B (Wagner and Fischer, 1974). Formally, the trace is defined as any set of ordered pairs of integers (i,j) satisfying – (1) $1 \le i \le n$ and $1 \le j \le m$; (2) for any two distinct pairs (i_1,j_1) and (i_2,j_2) in T, (a) $i_1 \ne i_2$ and $j_1 \ne j_2$; (b) $i_1 < i_2$ iff $j_1 < j_2$ (Wagner and Fischer, 1974). In (1), the pair (i,j) describes a pair between $a_i \in A$ and $b_j \in B$. I and J are the sets of positions in A and B respectively paired with empty string Λ. For the exemplary trace shown in Fig. 3a, the values of (i,j), I and J are {(1,1),(2,2),(3,3)}, $\{\phi\}$ and $\{\phi\}$, where $\{\phi\}$ is an empty set, while for the trace in Fig. 3b the respective values are {(1,1),(2,3)}, {3} and {2}. The γ in (1) is a cost function which assigns, to each pair $a \to b$, a non-negative real number satisfying following properties: (1) $\gamma(a \to a) = 0$; (2) $\gamma(a \to b) + \gamma(b \to c) \ge \gamma(a \to c)$. The optimal cost C_T^* to the function in (1) is obtained using two steps – (1) initialization, and (2) forward pass, as defined below. This is done by computing a matrix $D_{n\times m}$ such that its (i,j)th element $d_{i,j} = \min\ C_T(A(1:i), B(1:j))$; hence, $C_T^* = d_{n,m}$. The trace corresponding to the cost C_T^* is obtained using back-tracking algorithm from $i = n$, $j = m$ to $i = 1$, $j = 1$ implementing the back-tracking step defined below.

1. Initialization: $d_{0,0} = 0$; $d_{i,0} = \sum_{r=1}^{i} \gamma(a_r \to \Lambda)$ and $d_{0,j} = \sum_{r=1}^{j} \gamma(\Lambda \to b_r)\ \forall\ 1 \le i < n$ and $1 \le j \le m$.
2. Forward pass: $z_1 = d_{i-1,j-1} + \gamma(a_i \to b_j)$, $z_2 = d_{i-1,j} + \gamma(a_i \to \Lambda)$, $z_3 = d_{i,j-1} + \gamma(\Lambda \to b_j)$; $d_{i,j} = \min_{1\le l\le 3} z_l$.
3. Back-tracking: $x = \text{argmin}_{1\le l\le 3}\, z_l$; $if\ x = 1$ or $2 \implies i = i-1$ and $if\ x = 1$ or $3 \implies j = j-1$.

In WFA, $\gamma(a \to b) = 0$ if $a = b$, 1 otherwise. We observe that, with the cost function, γ, the optimal trace could pair a nucleus in A to a non-nucleus in B, as illustrated in Fig. 3a, where, nucleus '*ao*' in A pairs with non-nucleus '*r*' in B. This indicates that the trace does not obey the second and third constraints of the PSA. However, Fig. 3b shows a desirable trace, which satisfies all the PSA constraints.

2.4. Proposed weighted Wagner–Fisher algorithm

The modified cost function of trace T from A to B in the proposed wWFA is given as:

$$C_T(A,B,\alpha) = \sum_{(i,j)\in T} \gamma_\alpha(a_i \to b_j) + \sum_{i\in I} \gamma_\alpha(a_i \to \Lambda) + \sum_{j\in J} \gamma_\alpha(\Lambda \to b_j) \quad (2)$$

where, we define $\gamma_\alpha(a \to b) = \alpha_{ab}\gamma(a \to b)$, α_{ab} is a non-negative weight associated with the operation $a \to b$ and satisfies the property $\alpha_{ab} + \alpha_{bc} \ge \alpha_{ac}$. Thus, the proposed γ_α also satisfies the properties: (1) $\gamma_\alpha(a \to a) = 0$; (2) $\gamma_\alpha(a \to b) + \gamma_\alpha(b \to c) \ge \gamma_\alpha(a \to c)$. Further, the optimal trace to $C_T(A,B,\alpha)$ is obtained by replacing γ with γ_α in the three steps of WFA described in Section 2.3. It is easy to see that the solution of (2) is identical to that of (1) when all weights are one.

2.5. Proposed weights computation for wWFA

We modify wWFA in (2) to incorporate the constraints proposed in Section 2.2 and compute the weights for wWFA. We discuss the modifications and the weights computation for each constraint separately below.

Constraint 1: In order to ensure this constraint, it is necessary to estimate the trace that avoids pairing of two phonemes belonging to two sub-sequences in A and B representing two different words. Under such constraint on trace, from the work proposed by Wagner and Fischer (1974), we observe that the overall sentence based weighted cost function in (2) can be written as a sum of word based weighted cost functions as follows:

$$C_T^w(A,B,\alpha) = \sum_{k=1}^{n_w} C_T(A_k, B_k, \alpha) \quad (3)$$

where, A_k and B_k are the respective kth word sub-strings in A and B. n_w is the total number words in either A or B. Thus, solving the cost function in (2) for trace automatically ensure the first constraint. Further, we also observe that the trace, denoted as T_w, belonging to the optimal cost $C_T^{w^*}$ in (3) is obtained by concatenating the traces, $T_1, T_2, \ldots, T_{n_w}$, sequentially, where T_k is the trace corresponding to the optimal cost $C_T^*(A_k, B_k, \alpha)$ of the kth word.

Constraint 2: We split the second constraint into three sub-constraints based on the $\left|\mathscr{S}_{A_k}\right|$ and $\left|\mathscr{S}_{B_k}\right|$, where $\mathscr{S}_{A_k}$ and $\mathscr{S}_{B_k}$ are the sets of syllable nuclei indices in A_k and B_k respectively and $|.|$ indicates the cardinality of the set. The three sub-constraints are given as:

1. $\left|\mathscr{S}_{A_k}\right| < \left|\mathscr{S}_{B_k}\right| \Rightarrow \alpha_{a_i\Lambda}, \alpha_{a_i b_j} = \infty\ \forall\ i \in \mathscr{S}_{A_k},\ j \notin \mathscr{S}_{B_k}$. $\alpha_{a_i\Lambda}$, $\alpha_{a_i b_j}$ are denoted by $\alpha_{i\Lambda}$, α_{ij} for brevity, from now onward.
2. $\left|\mathscr{S}_{A_k}\right| > \left|\mathscr{S}_{B_k}\right| \Rightarrow \alpha_{\Lambda j}, \alpha_{ij} = \infty\ \forall\ i \notin \mathscr{S}_{A_k},\ j \in \mathscr{S}_{B_k}$
3. $\left|\mathscr{S}_{A_k}\right| = \left|\mathscr{S}_{B_k}\right| \Rightarrow$ both first and second sub-constraints are satisfied.

The first sub-constraint is active when $N = \left|\mathscr{S}_{A_k}\right|$ and it ensures that a syllable nucleus in A_k is neither deleted nor paired with a phoneme other than syllable nucleus (denoted by POSN) in B_k. If it does so, the cost becomes ∞; hence, a finite cost can be obtained only when all syllable nuclei in A_k are paired with one of the syllable nuclei in B_k. Similarly, the second sub-constraint ensures that all the syllable nuclei in B_k are paired with one of the syllable nuclei in A_k and the third

sub-constraint ensures that no syllable nucleus in A_k or B_k is paired with a POSN.

Constraint 3: Let i_p and j_p be the primary stress label indices in A_k and B_k respectively. Then, in order to ensure this constraint, the syllable nucleus at i_p and that at j_p should be paired with neither Λ nor any other nucleus of syllable and a POSN. Hence $\alpha_{\Lambda j_p}, \alpha_{i_p \Lambda} = \infty$, $\alpha_{i_p j} = \infty \; \forall \; j \neq j_p$ and $\alpha_{i j_p} = \infty \; \forall \; i \neq i_p$. However, j_p is unknown but we assume that it can be estimated as follows – (1) consider jth element in $\mathscr{S}_{B_k}$ as having primary stress label and compute the cost $C_T(A_k, B_k, \alpha)$; (2) consider the j that results in the minimum $C_T(A_k, B_k, \alpha)$ as j_p.

The remaining weights are considered as 1, when those are not assigned after incorporating the second and third constraints.

2.6. Proposed PSA algorithm

Algorithm 1 shows the steps for PSA using proposed wWFA considering the weights described in Section 2.5.

Algorithm 1 PSA using proposed wWFA. Input: $A_k = \{A_k(1), A_k(2), \ldots, A_k(|A_k|)\}$, $B_k = \{B_k(1), B_k(2), \ldots, B_k(|B_k|)\}$; $\forall 1 \leq k \leq n_w$ and output: $T_w = \{T_1, T_2, \ldots, T_{n_w}\}$

Initialization: $T_w \leftarrow \{\emptyset\}$
for each word k from 1 to n_w **do**
  Initialization: $\alpha_{i_p \Lambda} = \infty, \mathscr{S}_{A_k}, \mathscr{S}_{B_k}$
  for each frame $j_p \in \mathscr{S}_{B_k}$ **do**
    Weight initialization: (1) $\alpha_{\Lambda j_p} = \infty$,
    (2) $\alpha_{i_p j} = \infty; 1 \leq j \leq |B_k| \;\; \& \;\; j \neq j_p$,
    (3) $\alpha_{i j_p} = \infty; 1 \leq i \leq |A_k| \;\; \& \;\; i \neq i_p$.
    if $|\mathscr{S}_{A_k}| \leq |\mathscr{S}_{B_k}|$ **then**
    $\alpha_{i\Lambda} = \infty; i \in \mathscr{S}_{A_k}$
    $\alpha_{ij} = \infty; i \in \mathscr{S}_{A_k}$ and $1 \leq j \leq |B_k| \;\; \& \;\; j \notin \mathscr{S}_{B_k}$
    end if
    if $|\mathscr{S}_{A_k}| \geq |\mathscr{S}_{B_k}|$ **then**
    $\alpha_{\Lambda j} = \infty; j \in \mathscr{S}_{B_k}$
    $\alpha_{ij} = \infty; j \in \mathscr{S}_{B_k}$ and $1 \leq i \leq |A_k| \;\; \& \;\; i \notin \mathscr{S}_{A_k}$
    end if
    Cost computation: $d_{0,0} = 0$;
    $d_{i,0} = d_{i-1,0} + \alpha_{i\Lambda}.\gamma(A_k(i) \rightarrow \Lambda); 1 \leq i \leq |A_k|$
    $d_{0,j} = d_{0,j-1} + \alpha_{\Lambda j}.\gamma(\Lambda \rightarrow B_k(j)); 1 \leq j \leq |B_k|$
    for each i from 1 to $|A_k|$ **do**
      for each j from 1 to $|B_k|$ **do**
        $z_1 = d_{i-1,j-1} + \alpha_{ij}.\gamma(A_k(i) \rightarrow B_k(j))$
        $z_2 = d_{i-1,j} + \alpha_{i\Lambda}.\gamma(A_k(i) \rightarrow \Lambda)$
        $z_3 = d_{i,j-1} + \alpha_{\Lambda j}.\gamma(\Lambda \rightarrow B_k(j))$
        $d_{i,j} = \min_{1 \leq l \leq 3} z_l$; $L(i,j) = \operatorname{argmin}_{1 \leq l \leq 3} z_l$
      end for
    end for
    $\mathscr{D}(j_p) = d_{|A_k|,|B_k|}$; $\mathscr{L}(j_p) = L$
  end for
  Back tracking: $i = |A_k|$; $j = |B_k|$; $T_k \leftarrow \{i, j\}$
  $\eta = \operatorname{argmin}_{j_p} \mathscr{D}(j_p)$; $B_{opt} = \mathscr{L}(\eta)$
  while $i \neq 0 \;\; \& \;\; j \neq 0$ **do**
    if $B_{opt}(i,j) == 1$ **then** $i = i - 1$; $j = j - 1$;
    else if $B_{opt}(i,j) == 2$ **then** $i = i - 1$;
    else $j = j - 1$;
    end if
    $T_k \leftarrow T_k \cup \{i, j\}$
  end while
  $T_w \leftarrow T_w \cup T_k$
end for

3. Database

We use ISLE (Menzel et al., 2000) corpus in all our experiments in this work. The corpus contains utterances from 46 non-native speakers (23 German (GER) and 23 Italian (ITA)) learning English. Each speaker uttered approximately 160 sentences. For each utterance in the data, phoneme transcriptions were available, which were obtained from forced-alignment with GMM-HMM based acoustic models learnt from ISLE data using HTK toolkit (Young et al., 2002). We refer to these phoneme transcriptions and the GMM-HMM based model as ISLE_GMM-estimated phoneme sequences and ISLE_GMM respectively. Following this, the ISLE_GMM-estimated phoneme transcriptions were corrected manually to reflect the speakers' pronunciation by a team of five linguists. In addition to these, in the data, the stress labels on the syllable nuclei were available for both ISLE_GMM-estimated and manually corrected phoneme sequences, which were obtained manually from the same team of linguists by assuring only one stressed syllable, referred to as primary stress, in each word. Also, we refer to the annotated stress labels on the syllable nuclei in the manually corrected data as existing labels. Further, we convert both the manually corrected and ISLE_GMM-estimated phoneme transcriptions to syllable transcriptions using P2TK syllabifier (Tauberer, 2018) and obtain their respective time-aligned boundaries using phoneme specific time-aligned boundaries. We refer to the respective syllable transcriptions and its time-aligned boundaries together as ISLE_GMM-estimated and manually corrected syllable data, respectively. Further the manually corrected syllable data is referred to as existing syllable data as it contains existing labels.

4. Experiments and results

The stress detection performance depends on the quality of the labels obtained from the proposed stress label assignment. Thus, we discuss performance of the stress label assignment task followed by its effectiveness in the stress detection task. For both the tasks, all the weights in the proposed wWFA method are set to one except the weights, that satisfies the constraints 2 & 3, which are set to ∞.

4.1. Stress label assignment

4.1.1. Experimental setup

We consider unweighted accuracy (Tepperman and Narayanan, 2005; Yarra et al., 2017) and F-score as objective measures for evaluating the proposed stress label assignment approach. For comparison, we implement a baseline scheme, referred to as WFA-baseline, as follows: (1) compute a trace using WFA (2) considering the paired phonemes in the trace, estimate the stress labels for each syllable nuclei. In this process, the syllable nuclei are marked as unstressed when they do not obtain any stress label. Table 1 summarizes the syllable data and annotated labels used for the experiments. We consider the existing labels to assign onto the ISLE_GMM-estimated syllable data using the existing and ISLE_GMM-estimated phoneme sequences. In order to evaluate the assigned labels, we consider the annotated stress labels on ISLE_GMM-estimated syllables as the ground truth.

4.1.2. Results and discussions

We estimate the stress labels using the proposed stress label assignment and the WFA-baseline, from which, the respective unweighted accuracies (F-score in brackets) are shown in column 2 and 3 of Table 2. Higher accuracy and F-score with the proposed method indicates the effectiveness of the proposed stress label alignment method compared to the baseline scheme. We analyze the effectiveness of each of the proposed constraints and their combinations in the label assignment task. The values in column 4–7 in Table 2 shows the accuracies and F-scores obtained separately with the proposed label assignment considering constraint 1, constraint 2, constraint 1 & 2 and constraint 1

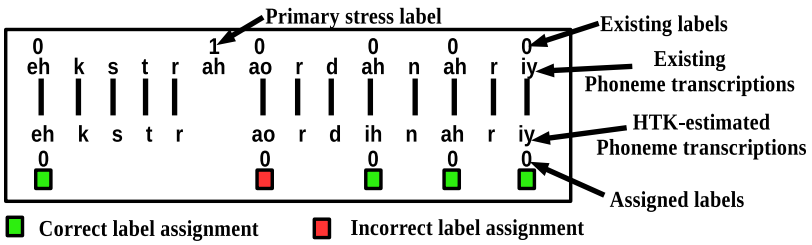

Fig. 4. An exemplary trace that results in incorrectly assigned labels on the ISLE_GMM-estimated phonetic transcriptions considering only constraint 1 in the proposed label assignment approach.

Table 1
Syllable data and annotated labels used (indicated by ✓) for assignment and evaluation in the stress label assignment experiments.

		Assignment		Evaluation
		Input	Output	
Existing (manually corrected)	Syllable data	✓		
	Annotated labels	✓		
ISLE_GMM	Syllable data	✓		
	Assigned labels		✓	
	Annotated labels			✓

Table 2
The label assignment accuracies and F-scores obtained separately WFA-baseline, the proposed approach and the proposed approach with constraint 1, constraint 2, constraint 1 & 2 and constraint 1 & 3. It is required to consider constraint 3 in conjunction with constraint 1; hence, the performance with constraint 3 and constraint 2 & 3 are not reported.

	WFA	Proposed	Proposed with constraint			
			1 only	2 only	1&2 only	1&3 only
Accuracy	97.92%	100%	98.12%	99.22%	99.66%	99.61%
F-score	0.9734	1	0.9759	0.9898	0.9956	0.9949

& 3 respectively. In the table, the performance with constraint 3 and constraint 2 & 3 are not reported, since, constraint 3 by itself cannot be used as it requires to be in conjunction with constraint 1. From the table, it is observed that all the accuracies and F-scores obtained with different constraints considered in the proposed approach are higher than the WFA-baseline accuracy and F-score (97.92% and 0.9734). This suggests the significance of the proposed constraints in the label assignment task. It is also observed that the accuracies and F-scores obtained by considering constraints 2 and 1 & 2 are higher than those with constraint 1 and 1 & 3 respectively. This indicates that constraint 2 is the most critical one among all the proposed constraints. It is also interesting to observe that none of the accuracies and F-scores are 100% and 1, which is achieved with the proposed method only when all three constraints are considered together. This indicates that using all three constraints together is required to achieve the highest performance in the proposed label assignment scheme.

Fig. 4 shows an exemplary trace obtained using the proposed label assignment considering only constraint 1. From the figure, it is observed that the labels are assigned incorrectly on the ISLE_GMM-estimated phonetic transcriptions. This could be because the syllable nuclei belonging to the primary stress label in the existing phoneme transcriptions is not paired with any syllable nuclei in the ISLE_GMM-estimated phoneme transcriptions. However, when all three constraints are considered in the proposed label assignment task, it results a trace that corrects above label assignment error. This suggests the significance of the three constraints proposed in this work.

Further, we analyze, using Table 3, the percentage of words that have incorrectly been assigned labels by WFA-baseline separately based on syllable count in a word. From the table, it is observed that the highest incorrect assignment happens in the six syllable words followed by three syllable words, while the least incorrect assignment

Table 3
The percentage of words with incorrectly assigned labels by WFA-baseline separately based on syllable count in a word.

	Syllable count in a word					
	1	2	3	4	5	6
WFA-baseline	0.71	0.76	1.39	0.89	0.00	5.56

is found in the five syllable words followed by one syllable words. This indicates that there is no trend in the assignment errors made by the WFA-baseline based on their exact syllable count. However, assignment errors of 0.74% and 1.96% are found when averaged across the words containing one to two syllables and three to six syllables, respectively. This indicates that, on average, the WFA-baseline makes more assignment errors when the words have relatively higher number of syllables. This suggests that the baseline scheme tends to make more errors in pairing of syllable nuclei in the existing syllable data with those in the ISLE_GMM-estimated syllable data when the number of syllables is large. However, the proposed label assignment approach eliminates such errors effectively.

4.2. Stress detection performance

4.2.1. Experimental setup

We consider unweighted accuracy (Tepperman and Narayanan, 2005; Yarra et al., 2017) and F-score as the objective measures for evaluation in the stress detection task too. Following the work by Yarra et al. (2017), we obtain the AFs for each syllable and estimate its respective stress label using an SVM classifier. The AFs are computed based on prominence measures such as intensity, duration and pitch along with sonority cues. These AFs have been shown to be effective compared to the features computed based on only prominence measures. The SVM classifier is implemented with RBF kernel with the complexity parameter (C) equal to 1.0 and with kernel coefficient (γ) equal to the inverse of the number of features. The SVM classifier is implemented using Scikit-learn (Pedregosa et al., 2011). In addition, we implement their post processing method, where, when the number of estimated stressed syllables in a word is different from one, the syllable with the highest decision score from the SVM classifier is declared as the stressed syllable.

Following their work, we use a speaker disjoint train and test set split considering 1st–12th & 1st–13th speakers' data for training and 13th–23rd & 14th–23rd speakers' data for testing for GER & ITA, respectively. We consider groups of the data from GER and ITA non-native speakers containing only polysyllabic words. This results in a total of 7586 & 7791 and 8586 & 4648 words in the train and test data, respectively, for GER & ITA speakers. There are a total of 3322 (4264) and 3723 (4863) stressed (unstressed) labels in the train and test data for GER speakers. Similarly, there are a total of 3411 (4380) and 2057 (2507) stressed (unstressed) labels in the train and test data for ITA speakers. In order to show the effectiveness of the proposed stress label assignment in the stress detection task, we use AFs from estimated syllable data in the test set to predict stress labels using

Table 4
Syllable data and annotated/assigned labels considered (indicated by ✓) for mismatched and matched stress detection experiments. The train and test splits for GER and ITA speakers are identical to those described in Para 2 of Section 4.2.1. The evaluation (Eval) measures are computed considering annotated labels on the subset of test splits for GER and ITA speakers as described in Para 2 of Section 4.2.2.

		Mismatched			Matched		
		Train	Test		Train	Test	
			Data	Eval		Data	Eval
Existing	Syllable data	✓					
	Annotated labels	✓		✓			✓
ISLE_GMM, FE_DNN,	Syllable data		✓		✓	✓	
LS_DNN, WSJ_DNN	Assigned labels				✓		

two classifiers trained separately as follows: (1) trained using AFs as similar to the work by Yarra et al. (2017) from the existing syllable data and existing labels, referred to mismatched train-test scenario, (2) AFs from the estimated syllable data and its respective labels obtained using the proposed stress label assignment, referred to as matched train-test scenario. Further, in order to show the effectiveness of the proposed method compared to WFA-baseline in the matched scenario, stress detection is performed using the labels obtained using WFA-baseline scheme.

In the experimentation, we consider four different sets of estimated syllable data obtained from forced-alignment considering one GMM-HMM and three different DNN-HMM based acoustic models. The GMM-HMM based estimated syllable data was available in the ISLE corpus i.e., ISLE_GMM-estimated syllable data. In order to obtain DNN-HMM based estimated syllable data, we use Kaldi speech recognition tool-kit (Povey et al., 2011) and a lexicon combining the following four lexicons – CMU (Weide, 1998), TIMIT (Zue et al., 1990), Beep (Robinson, 1996) and the lexicon used in preparing ISLE data. The phonemes in the combined lexicon are mapped to a set of 39 phonemes (Weide, 1998) following the phoneme mapping[1] available in the Kaldi tool-kit. Three DNN-HMM models are learnt by following the Daniel Povey's (Dan's Cosi, 2015) implementation (Povey et al., 2014) available in the Kaldi tool-kit using the three speech corpora respectively – (1) Fisher English (Cieri et al., 2004) (FE) (2) Libri-speech (Panayotov et al., 2015) (LS) (3) Wall street journal (Paul and Baker, 1992) (WSJ). We refer the DNN-HMM models based on FE, LS and WSJ as FE_DNN, LS_DNN and WSJ_DNN, respectively. We have not learnt a DNN-HMM acoustic model using ISLE corpus due to its limited amount of data. The labels for the three estimated syllable data is obtained by assigning the existing labels considering existing phoneme sequences and estimated phoneme sequences from the respective models. Table 4 summarizes the syllable data and annotated/assigned labels considered for the stress detection experiments for all the four types of the estimated syllable data under mismatched and matched conditions.

4.2.2. Results and discussions

In order to compare the accuracies and F-scores across all four sets of estimated syllable data, it is required to have identical number of syllables in the test set, while it is not guaranteed for all test cases. We summarize the mismatch between the syllables in the existing syllable data and those in each of the four sets of estimated syllable data in Table 5. The table shows the percentage of polysyllabic words in the test set that have the same and different number of syllables compared to those in the existing syllable data, for all four sets of estimated syllable data. From the table, it is observed that a significant percentage

[1] https://github.com/kaldi-asr/kaldi/blob/master/egs/timit/s5/conf/phones.60-48-39.map.

Table 5
Percentage of polysyllabic words that has difference of −3, −2, −1, 0 and 1 when the number of syllables in each of the four estimated syllable data subtracted from that in the existing syllable data.

Acoustic model for estimated data	−3	−2	−1	0	1
ISLE_GMM	0.00	0.29	14.98	84.68	0.05
FE_DNN	0.00	0.30	15.23	84.35	0.12
LS_DNN	0.00	0.41	18.11	81.40	0.08
WSJ_DNN	0.01	0.43	19.39	80.08	0.09

Table 6
Stress detection accuracies and F-scores obtained for the estimated syllable data under matched and mismatched conditions under WoPP & WPP for GER and ITA respectively. Also, when the stress detection is performed using AFs from the existing syllable data and label for both train and test conditions, stress detection accuracies (F-scores in brackets) are found to be 92.47% (0.8883) & 93.17% (0.8941) and 92.20% (0.8812) & 94.40% (0.9172) under WoPP & WPP for GER and ITA respectively.

		Mismatched		Matched			
				Proposed wWFA		WFA-baseline	
		WoPP	WPP	WoPP	WPP	WoPP	WPP
Accuracy							
ISLE_GMM	GER	77.80	84.75	90.19	91.39	89.99	91.21
	ITA	73.12	78.79	90.71	92.42	90.07	91.72
FE_DNN	GER	80.00	85.18	91.30	92.36	90.56	91.63
	ITA	78.51	83.72	92.63	94.50	91.80	93.53
LS_DNN	GER	76.37	83.04	91.20	92.81	90.55	92.24
	ITA	73.94	79.24	91.42	93.12	90.66	92.36
WSJ_DNN	GER	78.85	85.33	91.01	92.48	90.08	91.50
	ITA	75.46	80.84	91.51	93.52	90.64	92.64
F-score							
ISLE_GMM	GER	0.8100	0.8653	0.8657	0.8820	0.8611	0.8808
	ITA	0.7807	0.8303	0.8663	0.8873	0.8655	0.8879
FE_DNN	GER	0.8058	0.8609	0.8809	0.8966	0.8716	0.8877
	ITA	0.7960	0.8457	0.8898	0.9177	0.8801	0.9051
LS_DNN	GER	0.7741	0.8436	0.8810	0.9029	0.8734	0.8961
	ITA	0.7300	0.7949	0.8713	0.8977	0.8625	0.8880
WSJ_DNN	GER	0.7971	0.8574	0.8766	0.8986	0.8658	0.8868
	ITA	0.7511	0.8078	0.8740	0.9025	0.8639	0.8913

of the words has different number of syllables and that varies across the four sets. This indicates that the stress detection performance computed on the test set is not comparable directly across all four sets of estimated syllable data.

In order to circumvent this, we select a subset from the test data to ensure uniform test condition across all four sets considering existing syllable data as follows: (1) number of the syllables in a word in the existing syllable data is identical to that in the estimated syllable data from FE_DNN, LS_DNN, WSJ_DNN and ISLE_GMM (2) the existing labels on the syllables are identical to the assigned labels from the proposed wWFA based label assignment. This results in a subset of 5711 and 3941 words for GER and ITA speakers respectively. However, the training data varies across all four sets of estimated syllable data, in which, we consider all the syllables belonging to all the words in the entire train set.

Table 6 shows the accuracies and F-scores obtained using estimated syllable data from FE_DNN, LS_DNN, WSJ_DNN and ISLE_GMM. For each set of data, we compute the accuracies and F-scores with and without post processing (WPP and WoPP) under matched and mismatched scenarios. From the table, it is observed that the performance in the mismatched conditions are lower than those in the matched condition under WoPP and WPP for all four sets of estimated syllable data. This could be because of the fact (as observed in Table 5) that the variability in the number syllables could cause different time aligned boundaries between estimated syllable data and existing syllable data. Hence, the characteristics of AFs in the estimated syllable data would not be identical to those in the existing syllable data. Thus, the difference

in the characteristics of AFs could degrade the performance in stress detection task under mismatched scenario. This indicates that the stress detection performance degrades when the classifier is trained with existing syllable data and tested on estimated syllable data. Hence, in real test scenarios, it is useful to train a classifier for stress detection task using estimated syllable data and its respective labels obtained using the proposed stress label assignment.

It is also observed that the accuracies and F-scores under matched scenario are more when the estimated syllable data is obtained from three DNN-HMM based models (FE_DNN, LS_DNN and WSJ_DNN) compared to that from GMM-HMM based models (ISLE_GMM). This indicates that the features derived by considering DNN-HMM models perform better in stress detection task compared to those from GMM-HMM models. Further, it suggests that a better ASR model could result in better stress detection performance. It is interesting to observe that the accuracies and F-scores obtained with the proposed stress label assignment are higher than those obtained using the labels estimated with WFA-baseline for all four sets of estimated syllable data. Among all the four sets, the highest absolute improvements with the proposed wWFA are found to be 0.94% & 0.99% and 0.87% & 0.88% under WoPP & WPP on GER and ITA speakers, respectively, compared to WFA-baseline. We believe that these improvements are significant as those are comparable to the improvements (−1.28% & 1.72% and 1.16% & 3.09% under WoPP & WPP on GER and ITA speakers respectively) reported in the work by Yarra et al. (2017) on the same train and test splits. Also, the absolute improvements are comparable to the improvement of 1.4% reported in the work by Shahin et al. (2016) irrespective of speakers and WoPP when the baseline of single hidden layer multi-layer perceptron (MLP) is considered. These together suggest the benefit of the proposed stress label assignment algorithm.

Further, we perform stress detection similar to the work by Yarra et al. (2017) using AFs from existing syllable data and label for both train and test conditions. From this, we obtain accuracies (F-scores in brackets) of 92.47% (0.8863) & 93.17% (0.8941) and 92.20% (0.8812) & 94.40% (0.9172) under WoPP & WPP for GER and ITA, respectively. We compare these accuracies and F-scores with those in Table 6 in the matched scenario to examine the effectiveness of the AFs from the estimated syllable data. From the table, it is observed that a comparable accuracy (F-score) of 94.50% (0.9177) is achieved under WPP for ITA speakers using estimated syllable data from FE_DNN, where DNN-HMM based models are used. This indicates that the performances in stress detection task are comparable between the syllable data from DNN-HMM based models and manual annotation.

5. Conclusions

We propose a method to assign existing stress labels on the existing syllable data to an estimated syllable data obtained from forced-alignment to avoid time-consuming manual labeling. As there is no one-to-one correspondence between the estimated syllable data and existing labels due to mismatch between the number of syllables in them, we develop an algorithm by adding weights to the edit distances involved in Wagner–Fisher algorithm and compute those weights by defining a set of three constraints. Experiments on ISLE corpus show that the performance obtained on a test set for four different types of estimated syllable data are better when the assigned labels and estimated syllable data are used for training compared to those when existing labels and existing syllable data are used. Further investigations are required to analyze the benefit of the proposed method on different non-native corpora.

CRediT authorship contribution statement

Chiranjeevi Yarra: Methodology, Conceptualization, Validation, Formal analysis, Investigation, Writing – original draft. **Prasanta Kumar Ghosh:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Chandel, A., Parate, A., Madathingal, M., Pant, H., Rajput, N., Ikbal, S., Deshmukh, O., Verma, A., 2007. Sensei: Spoken language assessment for call center agents. In: IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). IEEE, pp. 711–716.

Cieri, C., Miller, D., Walker, K., 2004. The fisher corpus: a resource for the next generations of speech-to-text. In: 4th International Conference on Language Resources Evaluation, Vol. 4. pp. 69–71.

Cosi, P., 2015. A KALDI-DNN-based ASR system for Italian. In: International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–5.

Deshmukh, O.D., Verma, A., 2009. Nucleus-level clustering for word-independent syllable stress classification. Speech Commun. 51 (12), 1224–1233.

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., Precoda, K., 2015. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. Speech Commun. 69, 31–45.

Li, K., Zhang, S., Li, M., Lo, W.K., Meng, H.M., 2011. Prominence model for prosodic features in automatic lexical stress and pitch accent detection. In: Proceedings of Interspeech. pp. 2009–2012.

Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., Souter, C., 2000. The ISLE corpus of non-native spoken English. In: Proceedings of Language Resources and Evaluation Conference (LREC), Vol. 2. European Language Resources Association, pp. 957–964.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an ASR corpus based on public domain audio books. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5206–5210.

Paul, D.B., Baker, J.M., 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language. pp. 357–362.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., 2011. The kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE Signal Processing Society.

Povey, D., Zhang, X., Khudanpur, S., 2014. Parallel training of DNNs with natural gradient and parameter averaging. arXiv preprint arXiv:1410.7455.

Robinson, A., 1996. BEEP pronunciation dictionary. Retrieved from World Wide Web: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz.

Shahin, M.A., Ahmed, B., Ballard, K.J., 2014. Classification of lexical stress patterns using deep neural network architecture. In: Spoken Language Technology Workshop (SLT), 2014. IEEE, pp. 478–482.

Shahin, M., Gutierrez-Osuna, R., Ahmed, B., 2016. Classification of bisyllabic lexical stress patterns in disordered speech using deep learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6480–6484.

Tamburini, F., 2003. Prosodic prominence detection in speech. In: Seventh International Symposium on Signal Processing and its Applications, Vol. 1. IEEE, pp. 385–388.

Tauberer, J., 2018. P2TK automated syllabifier. Available at https://sourceforge.net/p/p2tk/code/HEAD/tree/python/syllabify/, Last Accessed on 14-03-2018.

Tepperman, J., Narayanan, S., 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). Citeseer, pp. 937–940.

Verma, A., Lal, K., Lo, Y.Y., Basak, J., 2006. Word independent model for syllable stress evaluation. In: International Conference on Acoustics Speech and Signal Processing (ICASSP), Vol. 1. IEEE, pp. 1237–1240.

Wagner, R.A., Fischer, M.J., 1974. The string-to-string correction problem. J. ACM 21 (1), 168–173.

Weide, R., 1998. The CMU Pronunciation Dictionary, Release 0.6. Carnegie Mellon University.

Yarra, C., Deshmukh, O.D., Ghosh, P.K., 2017. Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5845–5849.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., 2002. The HTK Book, Vol. 3. Cambridge University Engineering Department, p. 175.

Zhao, J., Yuan, H., Liu, J., Xia, S., 2011. Automatic lexical stress detection using acoustic features for computer assisted language learning. In: Proceedings of Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (ASC). pp. 247–251.

Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Commun. 9 (4), 351–356.